Performance Study of Phylogenetic Methods: (Unweighted) Quartet Methods and Neighbor-Joining

Katherine St. John

Dept. of Math & Computer Science, Lehman College and the Graduate Center, City U. of New York
E-mail: stjohn@lehman.cuny.edu

and

Tandy Warnow

Dept. of Computer Sciences, U. of Texas at Austin
E-mail: tandy@cs.utexas.edu

and

Bernard M.E. Moret

Dept. of Computer Science, U. of New Mexico
E-mail: moret@cs.unm.edu

and

Lisa Vawter

Aventis Pharmaceuticals
E-mail: lisa_vawter@aventis.com

We present the results of a large-scale experimental study of quartet-based methods (quartet cleaning and puzzling) for phylogeny reconstruction. Our experiments include a broad range of problem sizes and evolutionary rates, and were carefully designed to yield statistically robust results despite the size of the sample space. We measure outcomes in terms of numbers of edges of the true tree correctly inferred by each method (true positives). Our results indicate that these quartet-based methods are much less accurate than the simple and efficient method of neighbor-joining, particularly for data composed of short to medium length sequences. We support our experimental findings by theoretical results that suggest that quartet-cleaning methods are unlikely to yield accurate trees with less than exponentially long sequences. We suggest that a proposed reconstruction method should first be compared to the neighbor-joining method and further studied only if it offers a demonstrable practical advantage.

1. INTRODUCTION

Reconstructing the evolutionary history of a group of taxa is a major research thrust in computational biology. An evolutionary history not only gives relationships among taxa, but also an important tool for determining structural, physiological, and biochemical properties [5, 30]. Research on tree reconstruction has focused on reconstructing an evolutionary tree (phylogeny) under various optimization criteria. However, almost all optimization problems of interest to biologists are NP-hard (see [11] for a review), so most biologists use heuristic methods or surrogate optimization criteria.

A popular family of phylogenetic heuristics is based on *quartets*. A quartet is an unrooted binary tree for a quadruple of taxa. For most optimization problems, it is possible to determine the optimal tree on a set of four leaves by analyzing all three possible trees. Quartet-based methods compute a quartet under an optimization criterion for each set of four taxa and then combine the quartets to yield a tree on the full set of taxa. Because there are $\Theta(n^4)$ quartets, many quartet-based methods run in $\Omega(n^5)$ time, which is currently impractical for a hundred or more taxa.

How accurate are quartet-based methods? In biological applications, the true, historical tree cannot be ascertained exactly, which makes assessing the quality of reconstruction methods problematic (one exception is laboratory-created phylogenies for viruses and some bacteria, as illustrated in the study by Hillis *et al.* [10]). As a consequence, the method of choice for evaluating heuristics has been simulation [9]. In such a simulation, an ancestral biomolecular (DNA, RNA, or amino-acid) sequence is evolved along a "model tree," producing a synthetic set of biomolecular sequences at the leaves. Phylogenetic reconstruction methods are then assessed based upon how accurately they reconstruct the model tree (the "true" tree). Biologists typically evaluate performance according to the topological accuracy of the reconstructed unrooted tree, because the tree topology is interpreted as the order of past evolutionary events, that is, it yields the relationships among species, genes, or other taxa. (The reconstructed tree is typically unrooted, as determining a root is a very difficult problem of its own.) Topological accuracy is typically measured by the true positives, i.e., the percentage of edges of the true tree found in the reconstructed tree.

Among the distance-based methods (methods that transform input sequences into a distance matrix and then construct the tree from that distance matrix), none is more widely used by biologists than the *neighbor-joining (NJ)* method [27]. Not only is it quite fast ($O(n^3)$ for n taxa [29]), but experimental work has also shown that the trees NJ constructs are reasonably accurate, as long as the rate of evolution is neither too low nor too high. However, there is no comparative study of NJ and quartet-based methods.

We present the results of a detailed, large-scale experimental study of quartet-based methods and NJ under the Jukes-Cantor model of evolution [16]. Our results indicate that, under this model, NJ always outperforms the quartet-based

methods we examined, in terms of both accuracy and speed. We suggest that NJ, already the most popular distance-based method, should be used as a minimum standard in the assessment of phylogenetic methods: a proposed method should be compared with NJ and shown to be at least comparable in performance to NJ before it is studied in depth. We also present new theory about convergence rates of quartet-based methods which helps explain our observations.

2. TERMINOLOGY AND REVIEW

2.1. Simulations and the Jukes-Cantor Model

A *model tree* for sequence evolution is a pair, $(T, \{\lambda_e\})$, where T is a rooted unlabeled tree and, for each edge e of T, λ_e is the expectation of a Poisson distribution describing the number of changes at each site in the sequence along edge e. The *Jukes-Cantor* model [16] is the simplest Markov model of biomolecular sequence evolution. In that model, a DNA sequence (a string over the set of the four nucleotides: $\{A, C, T, G\}$) at the root evolves down a rooted binary tree. The sites (i.e., the positions within the sequences) evolve independently and identically, with equal probabilities of transition from one nucleotide to any other. A Jukes-Cantor model tree is a model tree in which the site changes take place according to the Jukes-Cantor model.

2.2. Measures of Accuracy

Let T be the true tree (that used in the model tree) and let T' be a tree produced by a reconstruction method, with both T and T' leaf-labelled by a set S of taxa. The edges of T' are often called the *reconstructed edges* since the method is trying to reconstruct the original tree. The true and inferred trees are compared only with respect to their underlying unrooted versions, in part because the reconstructed tree is typically unrooted and in part because the topological structure of a tree does not depend on the location of its root; therefore, in the remainder of this discussion, all trees are assumed to be unrooted.

For each edge $e \in E(T)$, we define the bipartition π_e induced on S by the deletion of the edge e from T. The bipartition π_e can also be written as the *split*, $\{A \mid B\}$, where A consists of all the leaves (that is, elements of S) on one side of the bipartition and where B is just $S - A$. (These definitions apply to any tree on the set S of leaves.) Methods for reconstructing trees are evaluated according to the degree of topological accuracy obtained, by comparing the sets of splits or bipartitions of the two trees. The *true positives* are the edges $e \in E(T)$ whose split also occurs in the splits of T'. Figure 1 illustrates this concept (note that the trees are drawn as rooted, but are compared only with respect to their unrooted versions).

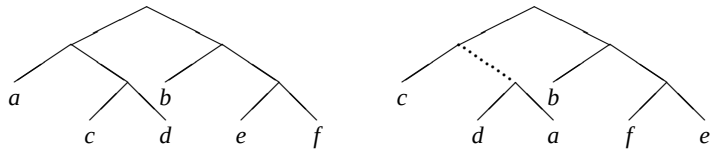


FIG. 1. True Positives: The true tree is on the left. In the tree on the right, the true positives (with respect to the tree on the left) are indicated by solid lines. The other edge of the tree is indicated by dotted lines.

2.3. Statistical Performance Issues

Under the Jukes-Cantor model, a method M is *statistically consistent* if, for every model tree $(T, \{\lambda_e\})$ and every $\varepsilon > 0$, there is a sequence length k (which may depend on M, T, $\{\lambda_e\}$, and ε) such that M recovers the unrooted version of T with probability at least $1 - \varepsilon$, when the method is given sequences (generated under the Jukes-Cantor model on $(T\{\lambda_e\})$) of length at least k.

The sequence length required by a method is a significant aspect of its performance, because real data sets are of limited length (typically bounded by a few hundred to a few thousand nucleotides). Computational requirements are also important, but it may be possible to wait longer or use more powerful machines, whereas it is not possible to get longer sequences than exist in nature. Consequently, experimental and analytical studies have attempted to bound the sequence lengths required by different phylogenetic methods. The rate at which a method converges to 100% accuracy as a function of the sequence length is called the *convergence rate*.

2.4. Neighbor-Joining

Neighbor-Joining (NJ) was formally described in 1987 [27] and has been a mainstay of phylogeny reconstruction among biologists ever since. NJ is a cubic-time distance-based algorithm that begins by creating a node for every taxa in the input set and making a list of those nodes. It proceeds by repeatedly pairing the two "closest" nodes from the list, adding a new node ("the parent") with edges to the selected pair, and replacing that pair with a new node. As this process progresses, the list of nodes available to pair shrinks and an acyclic graph is formed by the edges added with the new nodes. This process continues until the list of nodes is empty and a tree on all the nodes has been created. As with all distance-based phylogenetic reconstruction methods, the input to NJ is a dissimilarity matrix $\{d_{ij}\}$, where in practice d_{ij} is corrected, according to assumptions about the stochastic model of evolution underlying the data, in order to account for multiple hits. Even with this correction, however, it does not necessarily follow that two leaves i and j are siblings if d_{ij} is the minimum value (even if d is a tree metric). Therefore, NJ computes a secondary dissimilarity matrix $\{s_{ij}\}$, for which it does follow that i and j are siblings if s_{ij} is minimized and d is a tree metric. At each step, then, NJ chooses the pair of leaves with the smallest s_{ij} distance. NJ

eventually returns a binary (i.e. fully resolved) tree and is statistically consistent for the Jukes-Cantor model of evolution, and for any model for which statistically consistent distance corrections exist. (See [6], §7.3, for a very readable discussion of NJ, and a sketch of Atteson's proof [1] of its statistical consistency.)

2.5. Quartet-Based Methods

A *quartet* is an unrooted binary tree on four taxa. A quartet induces a unique bipartition of the four taxa and can be denoted by that bipartition. If the taxa are $\{a, b, c, d\} \subseteq S$, we can use $\{ab|cd\}$ to denote the quartet that pairs a with b and c with d (see Figure 2). A quartet $\{ab|cd\}$ *agrees* with a tree T if all four of its taxa are leaves of T and the path from a to b in T does not intersect the path from c to d in T. Equivalently, $\{ab|cd\}$ agrees with a tree if the homeomorphic subtree induced in T by the four taxa is the quartet itself. The quartet $\{ab|cd\}$ *is an error* with respect to the tree T if it does not agree with T. Figure 3 illustrates this idea. Let $Q(T) = \{\{ab|cd\} \mid a, b, c, d \in S$ and $\{ab|cd\}$ agrees with $T\}$. If $Q(T)$ denotes the set of all quartets that agree with T, then T is uniquely characterized by $Q(T)$ and can be reconstructed from T in polynomial time [7].

Quartet-based methods operate in two phases. In the first phase, they construct a set Q of quartets on the different sets of four taxa. A popular approach is to use maximum likelihood (ML), a computationally intensive but statistically sophisticated method [8, 22]. In the second phase, they combine these quartets into a tree on the entire set of taxa. In practice, not all quartets are accurately inferred, so it is necessary for quartet methods to handle incorrect quartets. Most optimization problems related to tree reconstruction from quartets are NP-hard. An example of this is the *Maximum Quartet Compatibility* problem [15], which seeks a tree T for a given set Q of quartets such that $|Q(T') \cap Q|$ is maximized.

The methods studied in this paper have no performance guarantees with respect to the *Maximum Quartet Compatibility* problem, although each of them is sta-

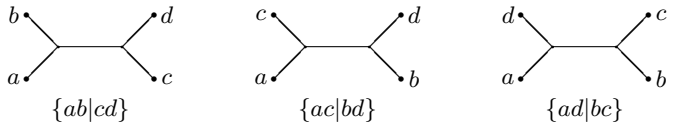
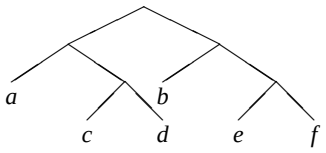


FIG. 2. The three possible quartets on four taxa $\{a, b, c, d\}$ and their bipartitions.



quartet $\{ab|cd\}$ is correct
quartet $\{ae|bc\}$ is an error

FIG. 3. For this model tree, the quartet $\{ab|cd\}$ *agrees* with the tree, while the quartet $\{ae|bc\}$ is *an error* with respect to the tree.

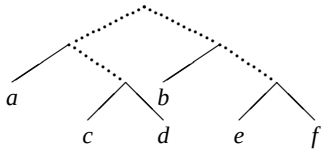


FIG. 4. The quartet $\{ce|df\}$ is in error around each of the dotted edges in the tree.

$\{ab|cd\}$
$\{ab|ce\}$
$\{ab|de\}$
$\{ae|cd\}$
$\{bc|de\}$

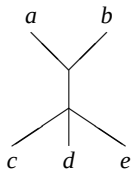


FIG. 5. Maximally resolved tree T' (right) of a given set of quartets (left).

tistically consistent under the Jukes-Cantor model of evolution. However, with the exception of Quartet Puzzling, all quartet methods we examine do provide guarantees about the edges of the true tree that they reconstruct. These guarantees are expressed in terms of *quartet errors around an edge*, a concept we now define.

Consider an edge e in the true tree T; its removal defines the split $\{A|B\}$ on the elements of S. Consider those sets of four leaves $\{a, a', b, b'\}$ with $\{a, a'\} \subseteq A$ and $\{b, b'\} \subseteq B$. Let t be a quartet on the leaves $\{a, a', b, b'\}$. Note that there are three such possibilities for t: $\{aa'|bb'\}$ (which agrees with the tree T) and two others: $\{ab'|a'b\}$ and $\{ab|a'b'\}$. The quartet t is said to be an *error around* e if t is not $\{aa'|bb'\}$ (i.e., if t does not agree with the tree T); see Figure 4. Similarly, if T' is a proposed tree, and Q is a set of quartets, then $t \in Q$ is an error around edge $e \in E(T')$ if $t = \{ab|a'b'\}$ or $t = \{ab'|a'b\}$, where π_e is the bipartition $\{A|B\}$.

Two of the methods we study, the Q^* method (also known as the Buneman method) and the Quartet Cleaning methods, can be described in terms of an explicit bound on the number of quartet errors around the edges they reconstruct. We begin with the Q^* method [3]. This method seeks the *maximally resolved* tree T' obeying $Q(T') \subseteq Q$. For example, Figure 5 contains a list of five quartets, Q. The only bipartition compatible with all quartets is $\{ab|cde\}$, so that the maximally resolved tree for these quartets, T', contains only a single nontrivial edge. The set of quartets induced by T', $Q(T')$, is a proper subset of the input set of quartets, Q:

$$Q(T') = \{\{ab|cd\}, \{ab|ce\}, \{ab|de\}\} \subset Q$$

Thus, by definition, there are *no quartet errors* around any edge in the tree T' with respect to the input set of quartets Q. This tree always exists, since the star

tree[1] trivially satisfies the constraint on any set of quartets. The Q^* tree is unique and can be constructed in polynomial time. By design, however, the Q^* method is conservative and generally produces very unresolved trees [13].

Quartet-Cleaning (QC) methods [2, 4, 15] have explicit bounds on the number of quartet errors around each reconstructed edge e. As with the Q^* method, we start with an input set of quartets Q. For each edge e in the true tree T, let q_e be the number of quartets that cross e—that is, all quartets $\{aa'|bb'\}$ where $a, a' \in A$ and $b, b' \in B$ and e induces the split $\{A|B\}$. The error bounds have the form $m\sqrt{q_e}$ where m is a small constant—the larger m, the larger the number of quartet errors around an edge that can be handled. The Q^* method can be viewed as a cleaning method in which $m = 0$. The *global cleaning* method sets $m = 1$, and the *local cleaning* method sets $m = \frac{1}{2}$. These methods are guaranteed to recover every edge of the true tree for which Q contains a small enough number of quartet errors. The *hypercleaning* method allows m to be an arbitrary integer and thus has the potential to recover more edges. However, its running time is very high—proportional to $n^7 \cdot m^{4m+2}$—so that it is impractical for m larger than 5.

The final quartet-based method we examined is the best known and the most frequently used by biologists [17, 20, 26]: the *Quartet-Puzzling (QP)* method [28]. This method computes quartets using an ML-based heuristic and then uses a greedy strategy to construct a tree on which many input quartets are in agreement. QP uses an arbitrary ordering of taxa, constructs the optimal quartet on the first four, then inserts each successive taxon in turn, attaching the new leaf to an edge of the current tree so as to optimize a quartet-based score. Because the input ordering of taxa is pertinent, QP uses a large number of random input orderings and computes the *majority consensus* of all trees found. (The majority consensus is the tree that contains all bipartitions that appear in more than half of the trees in the set and is a well-known consensus method among biologists.) Thus QP implicitly seeks to return a tree in which every edge is "well-supported," in the sense that each edge appears in more than half the trees obtained during the algorithm and has (presumably) many supporting quartets.

2.6. Previous Experimental Studies of Quartet Methods

Berry *et al.* conducted experimental studies of various QC methods [2, 4]. They evolved sequences on Kimura-Two-Parameter (K2P) model trees,[2] compared the quartets inferred by various methods with the quartets of the true tree, and determined which edges of the model tree could be reconstructed by their QC method. They varied evolutionary rates and sequence lengths, but only examined trees with 10 taxa. Their results showed that QC methods, especially hypercleaning, outperform the Q^* method with respect to true positives. By design, the QC methods

[1] The star tree on n leaves has $n + 1$ nodes and n edges and is composed of a central node to which all n leaves are directly connected.

[2] The K2P model [18] is a slight generalization of the Jukes-Cantor model in which the substitution rates among nucleotides are defined by two parameters, rather than just one.

cannot fail to recover an edge that is recovered by the Q^* method. So what is noteworthy in the experiments is that the QC methods *did* succeed in obtaining additional edges. Because the dataset sizes used in these experiments are quite small (only 10 taxa), these results may not generalize to larger numbers of taxa. Willson [31] used 12 taxa and the Jukes-Cantor model in conducting simulation studies to assess the accuracy of quartet inference by various methods, including (local) NJ, ML, maximum parsimony, and variants thereof. He found, as we did, that NJ (using distances corrected for the model of evolution) tended to return better quartets than ML under many conditions. Once again, however, the focus on a small fixed number of taxa limits the significance of the results. Finally, no comparison was made between QC methods and NJ or other tree reconstruction methods.

3. THEORETICAL BOUNDS ON THE CONVERGENCE RATES

We begin with the known upper bounds on the convergence rates of NJ and the Q^* method. Surprisingly, these are identical [1, 7], although experimental studies strongly suggest that NJ obtains accurate reconstructions of trees from shorter sequences than Q^* throughout the parameter space of Jukes-Cantor trees [13]. We then discuss upper bounds on quartet cleaning. Experimental results illustrating the tightness of the upper bounds can be found in [13, 14] and Section 5.7.

THEOREM 3.1. *Let $f, g, \varepsilon > 0$ be arbitrary constants with $f < g$. Denote by $Q^*(S)$ the tree reconstructed on S by the Q^* method and by NJ(S) the tree reconstructed by NJ. There is a constant $c > 0$ such that, for all Jukes-Cantor model trees $(T, \{\lambda_e\})$ on n leaves with $0 < f \leq \lambda_e \leq g < \infty$ for all $e \in E(T)$, and for a set S of sequences generated randomly on $(T, \{\lambda_e\})$,*

$$Pr[Q^*(S) = NJ(S) = T] > 1 - \varepsilon$$

if the sequence length exceeds $c \log n \cdot e^{O(g \cdot diam(T))}$, where $diam(T)$ is the largest number of edges among all paths in T. (Note that c depends on g and ε.)

Since the diameter of an n-leaf tree can be as much as $n - 1$ (and has expected value in $\Omega(\sqrt{n})$ for random trees [7]), Theorem 3.1 shows that the Q^* and NJ methods will converge from sequences that grow exponentially in n. While Theorem 3.1 provides only an upper bound, earlier experimental work shows that the Q^* method performs quite poorly when g and $diam(T)$ are both large [14], and that NJ is also affected, although less severely [13].

We now consider the convergence rates of the QC methods. The error bound used in QC methods is a multiple of $\sqrt{q_e}$, so that the ratio of permitted errors to the number of quartets around an edge is $m/\sqrt{q_e}$, where m is a constant depending on the choice of QC method. (Recall that we have $m = 1$ for the global cleaning method and $m = \frac{1}{2}$ for the local cleaning method.) Because q_e is $\Omega(n^2)$ and m

a small constant, this ratio rapidly approaches 0 as the number of taxa increases. For example, consider an edge in a 50-taxon tree producing a $20{:}30$ split. The number of quartets around this edge is $82,650$, so that the bound for local cleaning is only 144; hypercleaning with $m = 5$ brings this bound up to 1440. Thus, for 50 taxa, even hypercleaning has an error tolerance on some edges that is less than 2% of the total number of quartets for this edge.

The sensitivity of QC methods to errors suggests that, for large n, QC methods will be close in performance to the Q^* method. As soon as the number of errors around each edge exceeds the cleaning threshold, a QC method behaves identically to the Q^* method. As noted earlier, the convergence rate of the Q^* method is bounded from above by a function that grows exponentially in n, suggesting that the Q^* method might be impractical. If cleaning methods tend to perform only as well as the Q^* method for large n, then they will not scale well. In Section 5.7, we present experimental results that support this observation.

Consider therefore a hypothetical cleaning method we will call *HypoClean*. This method is guaranteed to recover an edge e if the number of quartet errors around e is at most one third of the quartets around the edge—a much more generous bound than that used in local cleaning or than that used by its authors in hypercleaning. In the following theorem, we establish a bound on the sequence length that suffices for *HypoClean* to be accurate on a random Jukes-Cantor tree.

We require the following lemma.

LEMMA 3.1. *The median diameter of all* $(2n-5)!!$ *unrooted, leaf-labelled, binary trees on* n *leaves is* $\Theta(\sqrt{n})$.

Proof. Penny and Steel [23] gave formulas for the distribution of interleaf distances in such trees under the assumption that all $(2n-5)!!$ such trees are equally likely, obtaining

$$\mu(D) = 2^{2n} / \binom{2n}{n}$$

and

$$\sigma^2(D) = 4n - 6 - \mu(D) - \mu^2(D)$$

Since any nondegenerate distribution must have its median within $[\mu-\sigma, \mu+\sigma]$, our conclusion follows. ∎

THEOREM 3.2. *Let* $f, g, \varepsilon > 0$ *be arbitrary constants with* $f < g$ *and denote by* $HC(S)$ *the tree reconstructed on* S *by the* HypoClean *method. Then there is a constant* c *such that, for Jukes-Cantor model trees,* $(T, \{\lambda_e\})$ *with* $0 < f \le \lambda_e \le g < \infty$ *for all* $e \in E(T)$*, and for a set* S *of* n *sequences generated randomly*

on T, we have $Pr[HC(S) = T] > 1 - \varepsilon$ whenever the sequence length exceeds $c \log n \cdot e^{O(g \cdot \sqrt{n})}$, where the constant c depends on g and ε.

Proof. Theorem 3.1 shows that quartets of low diameter are more easily reconstructed from short sequences than are quartets of high diameter. Assume that we can correctly reconstruct the "smallest-diameter" half of the quartets with high probability—we simply guess the remaining quartets. We will then correctly reconstruct $2/3$ of the quartets with high probability. What sequence length is required for this? Solving the smaller half of the quartets is no harder than solving the median-diameter quartets. By Theorem 3.1, this latter task is achieved with high probability when the sequence length is at least $O(c \log n \cdot e^{O(g \cdot md(T))})$, where $md(T)$ is the median diameter of T. By Lemma 3.1, this quantity is $\Theta(\sqrt{n})$. Therefore, the sequence length that suffices to reconstruct the true tree with high probability using the *HypoClean* method is $O(c \log n \cdot e^{O(g \cdot \sqrt{n})})$. ∎

We have established the same form of (exponential) upper bound on the sequence-length requirements of all cleaning methods. This upper bound suggests that cleaning methods may not scale well—if the upper bound is approached by any cleaning method, that method will require very long sequences to ensure high accuracy, yet such sequences may simply not be available.

4. EXPERIMENTAL DESIGN

4.1. Overview

We used Jukes-Cantor model trees with varying numbers of taxa and rates of evolution to generate a large number of synthetic datasets of varying lengths. For each dataset generated, we computed the NJ and QP trees on the entire dataset and two sets of quartets, one based upon (local) ML, Q_{ML}, and one based upon (local) NJ, Q_{NJ}. We then applied various cleaning methods to each of the sets Q_{ML} and Q_{NJ}. We use the terms "local" and "global" to distinguish between the local application of running a method on each set of four leaves to determine the quartet topology versus the global application of running the method on the complete set of leaves. We compared quartets of Q_{ML}, of Q_{NJ}, and of the reconstructed trees, as well as the reconstructed trees themselves, against the true tree for accuracy.

4.2. Model Trees

We randomly generated model tree topologies from the uniform distribution on binary leaf-labelled trees. For each edge of each tree topology, we generated a random number (from the uniform distribution) between 1 and 1000 and used that number as the "length" of the edge. We then scaled each such base model tree by a multiplicative factor, ranging from 10^{-7} to 10^{-3}. This process produces Jukes-Cantor trees with λ_e values ranging from a minimum of 10^{-7} to a maximum of 1. We generated random DNA sequences for the root and used the program

`Seq-Gen` [24] to evolve these sequences down the tree under the Jukes-Cantor model of evolution, thus producing sets of sequences at the leaves, our synthetic datasets.

4.3. Statistical Considerations

Because the number of distinct unrooted, leaf-labelled trees on n leaves is $(2n-5)!!$ and because our input space is further expanded by the choice of evolutionary rates, it is not possible to take a fair sample of the entire input space. In order to obtain statistically robust results, we followed the advice of McGeoch [19] and Moret [21] and used a number of *runs*, each composed of a number of *trials* (a trial is a single comparison), computed the mean outcome for each run, and studied the mean and standard deviation over the runs of these events. This approach is preferable to using the same total number of samples in a single run, because each of the runs is an independent pseudorandom stream. With this method, one can obtain estimates of the mean that are closely clustered around the true value, even if the pseudorandom generator is not perfect.

4.4. Parameter Space

A critical parameter of our study, one that has not been explored in most prior studies, is the number of input taxa. Previous experimental studies have often been limited to a small number of taxa due to computational problems. However, to resolve phylogenetic trees of interest to biologists, algorithms must scale reasonably, both in terms of topological accuracy and running time, to problems of the size that biologists typically study (20–200 taxa), as well as those they would like to address (a few hundred to several thousand taxa).

Because of the dedicated use of two multiprocessor clusters, we were able to run our test suite for 5, 10, 20, and 40 taxa (full quartet-based methods remain impractical, at least in terms of experimental studies, for large numbers of taxa). Our tests included a selection of eight expected evolutionary rates, from 5×10^{-5} to 5×10^{-1} per tree edge. For each evolutionary rate and problem size, we generated a total of 100 topologies, grouped into 10 runs of 10 trials. All tests were conducted for four sequence lengths: 500, 2,000, 8,000, and 32,000 (we note that sequence lengths above 1,000 are considered long and those above 5,000 extremely long; thus our study explores longer sequence lengths than are usually encountered in practice). In all, our study used 16,000 datasets and required many months of computation on the two clusters.

4.5. Algorithms

We tested four different phylogenetic reconstruction methods: NJ, local quartet-cleaning for quartets based on (local) NJ, local quartet-cleaning for quartets based on (local) ML, and QP. The code for QP is `TREE-PUZZLE`, available from their authors at `www.tree-puzzle.de`; we modified it only by removing its interactive interface. All other code is our own. For quartet-cleaning, our

accuracy measurements were made by counting the number of quartets that were in error around each edge. If the error was below the necessary threshold for the given method, then the edge was counted as being correctly reconstructed. For QP and NJ, we counted the number of true positives between the true tree and the tree method constructed. We ran all four algorithms sequentially on a single set of sequences for one trial, stored all data that was generated, then proceeded to the next trial, so that each of the algorithms was run on exactly the same data.

4.6. Measurements

Our focus in this study is the accuracy of solutions generated by the various tree reconstruction methods. Because most methods are time-consuming, the running time is briefly addressed; our aim was not to fine-tune implementations, but simply to obtain a rough idea of which methods can be run in a reasonable amount of time on a conventional machine for realistic datasets. We compare running times as gathered on our platforms, all of which are 450MHz Pentium III machines running Linux.

To assess topological accuracy, we measured the number of true positives (edges of the true tree that appear in the reconstructed tree). For cleaning methods, we measured these values before and after cleaning. For each run of 10 trials, we retained only the mean values. Our results are composed of the means for each set of 10 runs.

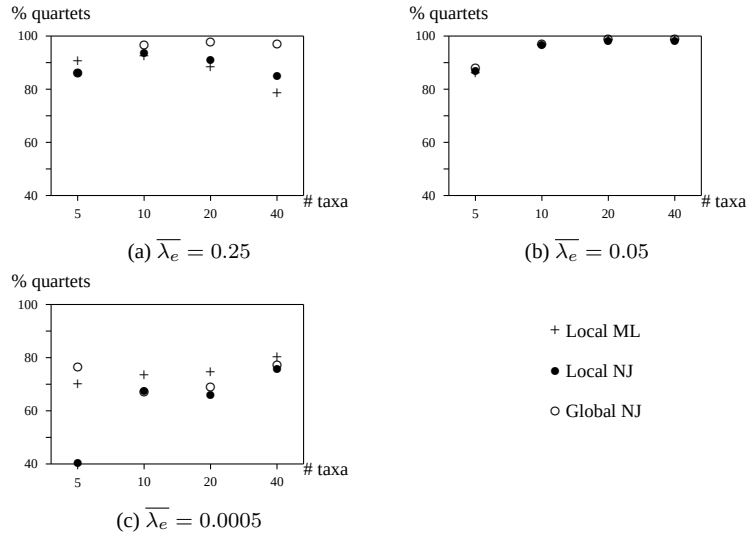
5. EXPERIMENTAL RESULTS

Except for runs on 5 taxa, the standard deviations we observed remained consistent at 1–2% of the mean; with 5 taxa, standard deviations were (as expected) larger, reaching 10–15% of the mean. In all of our figures, QCNJ and QCML denote quartet-cleaning of quartets derived by local NJ and by local ML, respectively.

5.1. Estimating Quartets

The technique used to construct the set Q of quartets provided to quartet-based methods can have a significant impact on the performance of these methods. The phylogenetics community has generally expected that (local) ML would produce more accurate quartets than (local) NJ. We therefore compared local ML and local NJ in terms of the quartet sets, Q_{NJ} and Q_{ML}, that they computed. As a reference point, we also examined how global NJ performed in terms of the trees it induced on each fourtuple of leaves from the global NJ tree. Figure 6 shows the proportion of true positives in each of the sets of quartets.

The relative performance of local NJ and local ML (NJ and ML applied to each quadruple of leaves to estimate the quartets) is interesting. At the highest rates of evolution (see Figure 6(a)), except for 5-taxon trees, local NJ slightly outperforms local ML, but this gap increases with increasing numbers of taxa. At the second



(a) $\overline{\lambda_e} = 0.25$ (b) $\overline{\lambda_e} = 0.05$

(c) $\overline{\lambda_e} = 0.0005$

FIG. 6. Percentages of quartets computed by local ML, local NJ and induced by global NJ that agree with the true tree for various numbers of taxa and a sequence length of 500. Here $\overline{\lambda_e}$ refers to the expected number of events on a random edge in the model tree.

highest rate of evolution (see Figure 6(b)), they are indistinguishable up to 40 taxa. However, at the lowest rate of evolution (see Figure 6(c)), local ML slightly outperforms local NJ, although the gap decreases with increasing numbers of taxa. Overall, while the choice of local NJ *vs.* local ML does influence the results, our data do not allow us to establish a preference for one over the other: neither ML nor NJ dominates the other in terms of accuracy, but each has a range in which it yields slightly better quartet estimations.

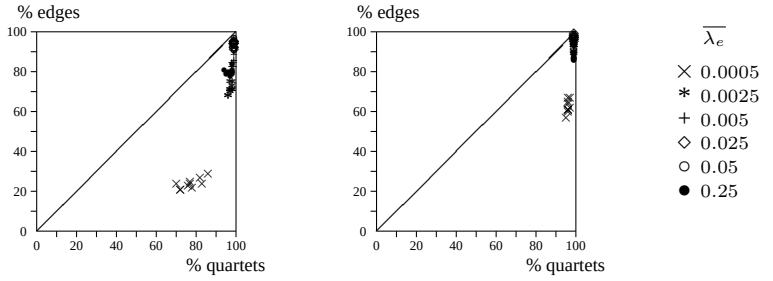
A comparison between these sets of quartets and the quartets obtained by using global NJ (i.e., the quartets induced by the NJ tree) is also interesting. At the lowest rate of evolution (Figure 6(c)), except for 5-taxon trees, local ML is superior to global NJ and both are superior to local NJ; however, the gap between the three ways of computing quartet trees narrows with increasing number of taxa. At the middle rate (Figure 6(b)), the methods are indistinguishable (up to 40 taxa), while at the highest rate (Figure 6(a)), global NJ is clearly superior, and the gap between global NJ as a quartet method and the two other quartet methods increases with increasing numbers of taxa. Thus, for high rates of evolution (and potentially for all large enough trees), the best quartet estimator may simply be global NJ—i.e., compute the NJ tree and use its quartets.

In terms of the quality of the quartets obtained, the best accuracy was obtained at the second highest rate of evolution. At the lowest rate of evolution, only 1 in 2000 sites changes on average around each edge, so that, for a sequence length of 500, roughly 25% of the edges have changes on them. Thus, although it may

be possible to guess an edge accurately, the best possible reconstruction at the lowest rate will only yield about 75% of the edges— approximately what the best performing method (local NJ) obtains. At the highest rate the accuracy starts to decrease with more than 10 taxa. A decrease in accuracy with increasing numbers of taxa for a fixed sequence length is predicted by theory (if only for information-theoretic reasons); hence, even for the lower rates of evolution, as the number of taxa increases, the accuracy of the quartet estimations should decrease.

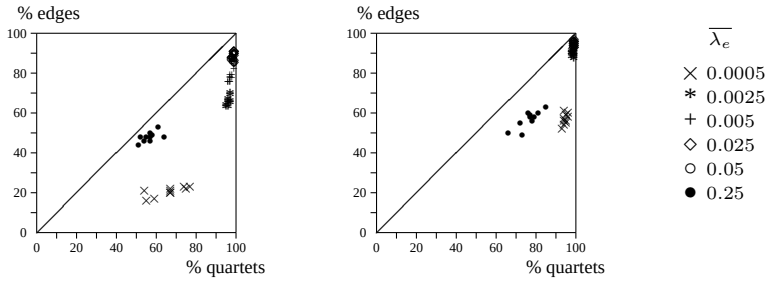
5.2. Two Measures of Accuracy: Quartets and Edges

Although the standard measure of accuracy is the number of true edges in the reconstructed tree, the percentage of correctly inferred quartets has also been used as a surrogate [4]. However, correlation between correct quartets and edges of the true tree returned by a method has not been shown. We address this issue by examining the performance of QP and global NJ with respect to both criteria. Figures 7 and 8 make it clear that edge accuracy is a more demanding criterion than quartet accuracy, and should therefore be used to assess performance of phylogenetic reconstruction methods. Both global NJ and QP can return trees with 80% of quartets correct, but only 20% of edges correct. Worse yet, both



(a) seq. length 500 *(b) seq. length 2000*

FIG. 7. Percent of true tree edges recovered by global NJ for various $\overline{\lambda_e}$ as a function of the percentage of correct induced quartets for 40 taxa and two sequence lengths.



(a) seq. length 500 *(b) seq. length 2000*

FIG. 8. Percent of true tree edges recovered by QP for various $\overline{\lambda_e}$ as a function of the percentage of correct induced quartets for 40 taxa and two sequence lengths.

methods, except when the percentage of correct quartets is close to 100%, can return fewer than 80% of the true tree edges (in the case of QP, some such trees had only 60% of the true tree edges). Because failure to obtain at least 90 or 95% of the edges can be unacceptable to systematists, quartet-based measures of accuracy are not acceptable surrogates for measures of accuracy based on true tree edges.

5.3. Sensitivity to Input Quality

Methods that operate by estimating quartets and then combining them into a single tree can be greatly affected by the quality of the input quartets. Figure 9 shows how QC methods respond to input quality. QC methods, as well as the other quartet methods we study, require a larger fraction of correct input quartets than the fraction of true tree edges that they return.

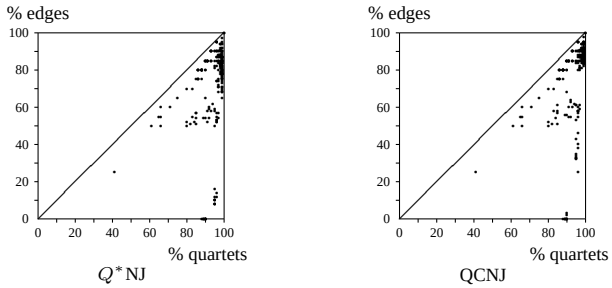


FIG. 9. Percentage of correct input NJ quartets *vs.* true tree edges for Q^*NJ and QCNJ for sequence length fixed at 2000, with each graph showing runs for all numbers of taxa and all average edge lengths.

5.4. Scaling of Methods with Increasing Numbers of Taxa

Theory predicts that the accuracy of methods will eventually degrade as the number of taxa increases while sequence length and average edge length (the expected number of changes for a random site on each edge) are held fixed. Figure 10 shows the edge accuracy achieved by all six methods as a function of the number of taxa for a sequence length of 500 and for three different average edge lengths. Figure 11 shows the same set of results for a sequence length of 2000. All methods decrease in accuracy as the number of taxa increases, even though both NJ and QP show an initial increase. QC provides a distinct improvement over the Q^* method, whether the quartets are computed using local ML or local NJ. QCML and QCNJ are very close in performance, although QCNJ slightly outperforms QCML; similarly Q^*NJ slightly outperforms Q^*ML. Of the five quartet methods, QP is the best throughout the range of parameters studied, but global NJ completely dominates it (and the other quartet methods we study).

5.5. A Comparison Between Q^* and QC

QC can be seen as an improvement to the Q^* method, because Q^* does not permit errors around any reconstructed edges, while QC reconstructs every edge around which there is a bounded number of errors. In Figures 10 and 11, we showed performance for different rates of evolution as the number of taxa varies, which gives evidence that QC methods return additional true edges under many conditions. In Figure 12, we explore the relative improvement in edge recovery obtained on local NJ or local ML quartets by using a QC method instead of the Q^* method. Curiously, the improvement obtained in terms of quartet accuracy is less satisfactory, never averaging more than 2% for low rates of evolution and for large number of taxa at high rates of evolution. QC provides the largest improvement when almost all input quartets are correct; indeed, this is what the theory about QC suggests. In particular, the most improvement occurs at a high rate of evolution—not our highest rate, but our second highest rate, when the error rate in input quartets is also lowest.

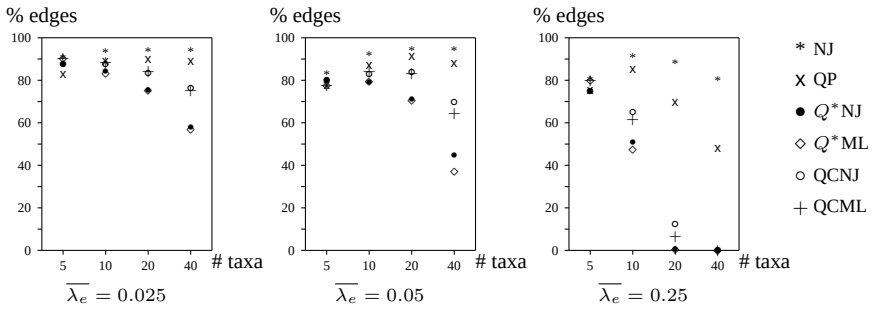


FIG. 10. Percentage of edges correct *vs.* number of taxa for sequences of length 500 and various $\overline{\lambda_e}$.

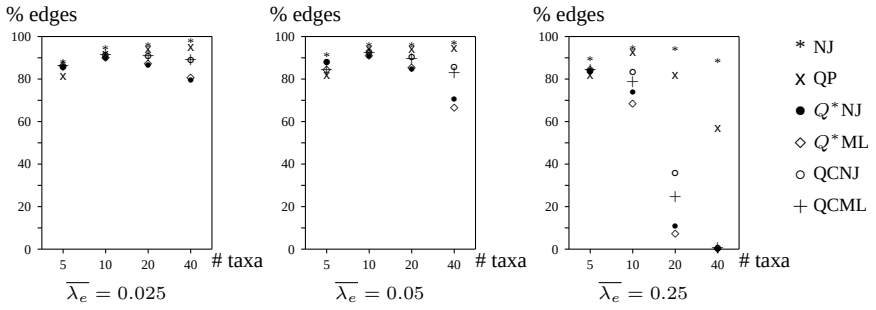
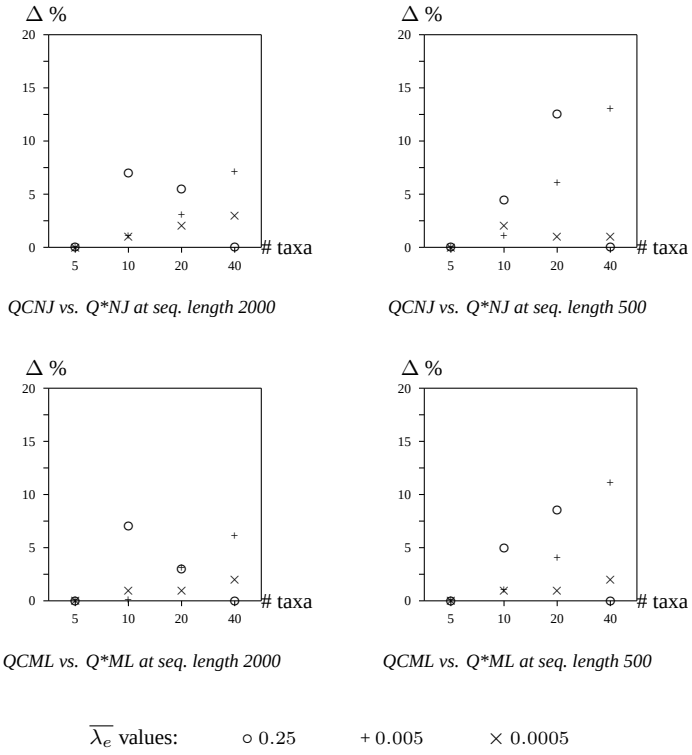


FIG. 11. Percentage of edges correct *vs.* number of taxa for sequences of length 2000 and various $\overline{\lambda_e}$.



*QCNJ vs. Q*NJ at seq. length 2000* *QCNJ vs. Q*NJ at seq. length 500*

*QCML vs. Q*ML at seq. length 2000* *QCML vs. Q*ML at seq. length 500*

FIG. 12. QC *vs.* Q^*: cleaning-induced improvement for NJ and ML in the percentage of tree edges that agree with the true tree. Δ is the additional percentage of true tree edges obtained by using quartet cleaning.

5.6. The Effects of Sequence Length

Although sequence length and rate of evolution have a strong effect on the absolute performance of phylogenetic methods, the relative ordering of NJ, QP, and QCNJ is almost constant throughout our experiments: NJ is the best followed by QP, and then by QCNJ. Figure 13 presents data for 40 taxa at three different rates of evolution, for sequence lengths varying from 500 (a typical length) up to 32,000 (a quite large length). Note that all methods increase in accuracy with increasing sequence length (as expected since all methods are statistically consistent under the Jukes-Cantor model).

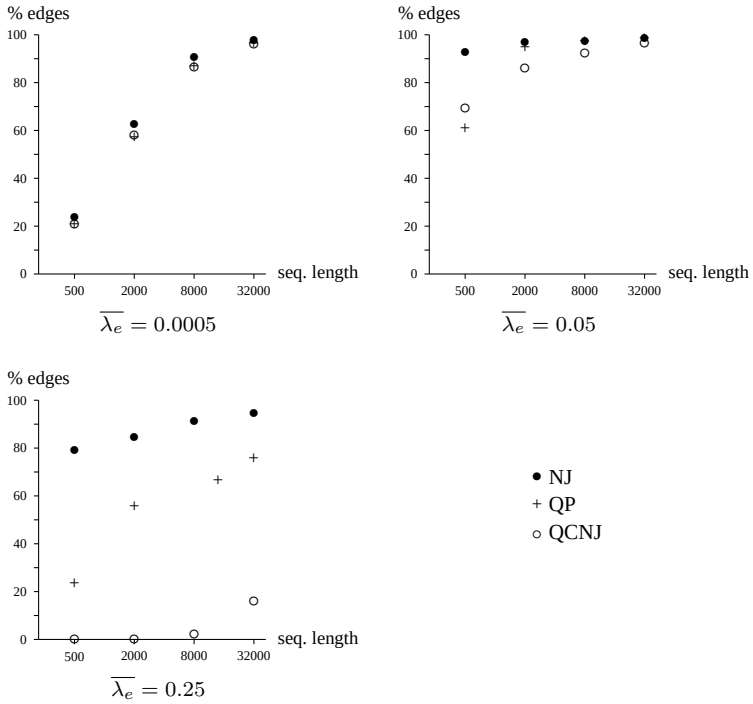


FIG. 13. Accuracy of various methods as a function of sequence length for 40 taxa.

5.7. Experimental Bounds on Sequence Length

Theorem 3.1 provides only an upper bound on the sequence length sufficient for accurate reconstruction by the Q^* method; no theoretical lower bound is known for the necessary sequence length. Using the same experimental set-up as before, we measured the sequence length required to recover accurately all of the edges at least 90% of the time for global and local quartet cleaning (with neighbor-joining determining the topology of the quartets). We generated a tree (under the distributions described above), evolved sequences down the tree of length 500,

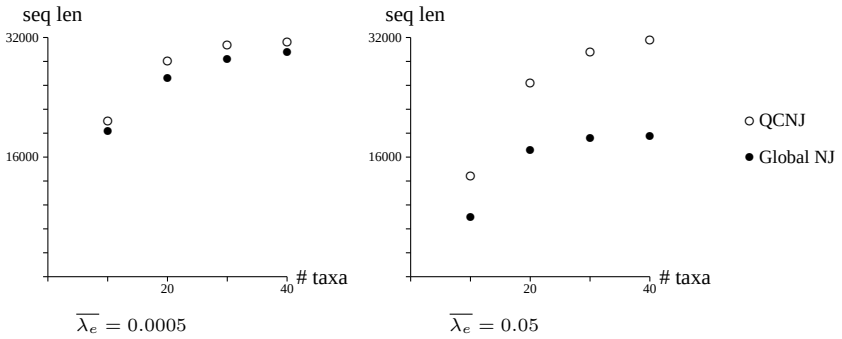


FIG. 14. The sequence length required to accurately reconstruct all of the edges 90% of the time. For low rates of evolution (left), QCNJ and NJ perform comparably. At higher rates of evolution (right), NJ consistently requires shorter sequences to reconstruct the true tree.

and used those sets of sequences as inputs to the methods. If a method fails to recover all of the edges of 9 out of 10 of the trees, the sequence length is increased by 500, and we repeat the reconstruction. We stop this process when a method has been successful, or the sequence length reaches 32,000 (a very large length). We note that, at all evolutionary rates, NJ outperforms QCNJ in the sequence lengths needed to reconstruct the edges 90% of the time (see Figure 14).

6. DISCUSSION

6.1. Quality of Quartets

The technique used to construct the set Q of input quartets provided to quartet-based methods can have a significant impact on the performance of these methods. The phylogenetics community has generally expected that local ML would produce more accurate quartets than other local quartet inference methods. However, in our studies, neither local ML nor local NJ dominates the other as a quartet estimator; instead, local ML outperforms local NJ only for the lowest rates of evolution, whereas local NJ clearly outperforms local ML for higher rates. Because our observations differ from the received wisdom in the field, we offer the following possible explanation. In earlier studies [12], the performance of local ML and local NJ as quartet estimators was studied by explicitly simulating evolution on 4-taxon trees. Here, we have simulated evolution on larger trees and then looked at the quartets defined by these larger trees. Good performance on quartets drawn from a large tree is not the same as good performance on quartets drawn from a very different sample space. While it is possible to sample 4-taxon model trees so as to produce the same kind of quartets we gave as input to our methods, the studies in [12] did not use such a sampling strategy.

6.2. Robustness of Quartet Methods to Quartet Errors

How robust are quartet-based methods with respect to errors in Q? The Q^* method is the least robust. QC methods provide some error tolerance, sufficient to recover additional true edges even under high rates of evolution and for moderate numbers of taxa. However, both of these methods are inferior to QP in terms of error tolerance, even though QP also fails to get a good estimation of the true tree when the input set of quartets has over 5% of errors (for $n = 40$). Finally, in our experiments, NJ was always at least as accurate as QP and nearly always much better. Thus, the reason quartet methods fail to recover good trees is *not* that the input distance matrix is too noisy for any method to recover a good estimate of the true tree.

6.3. Running Times

NJ was clearly the fastest method tested. QC and QP methods must compute all $\Theta(n^4)$ quartets and hence must take $\Omega(n^4)$ time. ML-based methods also construct quartets through expensive estimation methods, the running time of which increases sharply with increasing sequence length. Thus QCML and QP were by far the slowest of the methods tested, slow enough that running them on more than a hundred taxa appears infeasible at present. With default settings, QP takes more than 200 days of computation to analyze ten runs of ten trials each for a single set of parameters on 80 taxa with a sequence length of 500. In contrast, NJ dispatches the same analysis in about 30 minutes.

6.4. Comparison Between Methods

Our experiments clearly establish a linear order of accuracy for the methods we studied (except under very low rates of evolution): NJ (applied globally) is the preferred method, with QP second, the QC methods significantly behind QP, and the Q^* methods somewhat behind the QC methods. The particular technique used to infer quartets also has an influence on the quality of the trees obtained by the quartet methods, with QCNJ often better than QCML and Q^*NJ often better than Q^*ML (at least for large enough trees with moderate to high rates of evolution). Furthermore, global NJ requires significantly shorter sequences to reconstruct the trees than the quartet methods we studied.

7. CONCLUSIONS AND OPEN QUESTIONS

Why does global NJ outperform the quartet methods throughout the parameter space we examined (except on some 5-taxon trees)? As Figure 14 shows, the actual convergence rates for both global NJ and QCNJ appear much better than exponential, suggesting that our upper bounds on the convergence rate are loose for both cleaning methods and NJ. Yet the same figure also shows clearly that the convergence rate of NJ is much better than that of QCNJ. The sharp degradation in accuracy that we see in cleaning methods with increasing numbers of taxa suggests

that their convergence rate, while perhaps subexponential, is asymptotically poor. In contrast, global NJ (and, to a lesser extent, QP) degrades far more gracefully, and only when the rate of evolution is close to saturation. The good performance of QP as a quartet method does not seem to result from its use of ML-based quartets, since by that reasoning QCML should demonstrate a comparable improvement over QCNJ (which it does not). One reason for the better behavior of QP could be the manner in which it combines quartets. We suspect that the issue is partly that the Q^* and QC methods place too stringent a requirement on the edges; by comparison the QP method places no absolute restriction. Thus, we suspect that the ability of global NJ and QP to handle noisy input data lies in the specific techniques each uses to construct trees and the fact that neither places strict bounds on errors. This in itself may help explain why QP outperforms the other quartet methods we studied, but it does not explain why global NJ outperforms QP. We conjecture that methods which operate by combining quartets do not make use of all available information: we suggest that quartet-based methods may be impeded by their very structure, in having to decide the tree based on quadruples of taxa, without reference to the other taxa.

These observations suggest that quartet methods, if they are to be competitive with global NJ, need to be flexible in combining quartets into a single tree on the full set of taxa. Because of the lack of correlation between quartet accuracy and edge accuracy, seeking to solve the quartet compatibility problem may not produce the best trees either. Therefore, quartet methods with good performance (reaching or improving upon global NJ's performance) will require both more flexibility and greater sophistication than the current quartet methods.

Another experimental study of quartet-based methods [25] compared QP with variants of global NJ, ML, and maximum parsimony, on 12-taxon trees. They noted poor performance by QP with respect to the other methods studied, which they attribute to poor weighting of the quartets (pointing out how difficult it is to decide how much weight or confidence to give each quartet). This study, along with our results, suggest that some flexibility in weighting quartets could improve the accuracy of quartet-based methods. We conclude with the following comments about algorithm design and performance studies in phylogenetics. From the perspective of experimental performance studies and algorithm design, global NJ should be regarded as a universal lowest common denominator in phylogeny reconstruction algorithms. Its speed makes it easy to use under all circumstances; its topological accuracy makes it an acceptable starting point for tree reconstruction in biological practice. We suggest that a proposed method should be compared with NJ and abandoned if it does not offer a demonstrable advantage over NJ for substantial subproblem families.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees for their useful comments in improving the readability of this paper and suggesting improvements that led to the experiments in Section 5.7. The authors wish to thank the Department of Computer Science at University of New Mexico for hosting KS, TW, and LV in Summer 2000. KS would also like to thank the Department of Computer Sciences and the Texas Institute of Computational and Applied Mathematics at the University of Texas for hosting her during the 2000–2001 academic year.

This work was supported in part by National Science Foundation grants ACI 00-81404 (BM), CDA 95-03064 (BM), CCR 94-57800 (TW), DEB 01-20709 (BM and TW), DEB 01-211651 (KS), EIA 99-73874 (KS), EIA 01-13095 (BM), EIA 01-13654 (TW), EIA 01-21377 (BM), and EIA 01-21680 (TW), and by the David and Lucile Packard Foundation (TW).

REFERENCES

1. Atteson, K., "The performance of the neighbor-joining method of phylogeny reconstruction," *Algorithmica* **25**, 2/3 (1999), 251–278.
2. Berry, V., Bryant, D., Jiang, T., Kearney, P., Li, M., Wareham, T., and Zhang, H., "A practical algorithm for recovering the best supported edges of an evolutionary tree," *Proc. 11th Ann. ACM/SIAM Symp. Discrete Algs. SODA 2000*, SIAM Press (2000), 287–296.
3. Berry, V., and Gascuel, O., "Inferring evolutionary trees with strong combinatorial evidence," *Theor. Comp. Sci.* **240**, 2 (2000), 271–298.
4. Berry, V., Jiang, T., Kearney, P., Li, M., and Wareham, T., "Quartet cleaning: improved algorithms and simulations," *Proc. Europ. Symp. Algs. (ESA99)*, LNCS **1643**, 313–324.
5. Chambers, J.K., Macdonald, L.E., Sarau, H.M., Ames, R.S., Freeman, K., Foley, J.J., Zhu, Y., McLaughlin, M.M., Murdock, P., McMillan, L., Trill, J., Swift, A., Aiyar, N., Taylor, P., Vawter, L., Nahed, S., Szekeres, P., Hervieu, G., Scott, C., Watson, J.M., Murphy, A.J., Duzic, E., Klein, C., Bergsma, D.J., Wilson, S., and Livi, G.P., "A G protein-coupled receptor for Uridine 5'-diphosphoglucose (UDP-glucose)," *J. Biol. Chem.* **275** (2000), 10767–10771.
6. Durbin, R., Eddy, S., Krogh, A., and Mitchison, G., *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, Cambridge University Press, 1988.
7. Erdős, P.L., Steel, M., Székély, L., and Warnow, T., "A few logs suffice to build almost all trees—I," *Random Structures and Algorithms* **14** (1997), 153–184.
8. Felsenstein, J., "Evolutionary trees from DNA sequences: A maximum likelihood approach," *J. Mol. Evol.* **17** (1981): 368-376.
9. Hillis, D.M., "Approaches for assessing phylogenetic accuracy," *Syst. Biol.* **44** (1995), 3–19.
10. Hillis, D.M., Bull, J.J., White, M.E., Badgett, M.R., and Molineux, I.J., "Experimental phylogenies: generation of a known phylogeny," *Science* **255** (1992), 589–592.
11. Hillis, D.M., Moritz, C., and Mable, B. *Molecular Systematics*. Sinauer Pub., Boston, 1996.
12. Huelsenbeck, J., and Hillis, D. "Success of phylogenetic methods in the four-taxon case." *Syst. Bio.* **42 (3)**: (1993), 247–264.
13. Huson, D., Nettles, S., Rice, K., Warnow, T., and Yooseph, S., "The hybrid tree reconstruction method," *ACM J. Experimental Algorithms* **4**, 5 (1999), `www.jea.acm.org/1999/HusonHybrid/`.
14. Huson, D., Nettles, S., and Warnow, T., "Disk-covering, a fast converging method for phylogenetic tree reconstruction," *J. Comp. Biol.* **6** (1999), 369–386.
15. Jiang, T., Kearney, P.E., and Li. M., "A polynomial-time approximation scheme for inferring evolutionary trees from quartet topologies and its application," *SIAM J. Comput.* **30**, 6 (2001), 1942–1961.

16. Jukes, T.H., and Cantor, C. *Mammalian Protein Metabolism.* Academic Press, NY (1969), 21–132.

17. Keeling, P.J., Luker, M.A., and Palmer, J.D., "Evidence from beta-tubulin phylogeny that microsporidia evolved from within the Fungi," *Mol. Biol. & Evol.* **17** (2000), 23–31.

18. Kimura, M., "A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences," *J. Mol. Evol.* **16** (1980), 111–120.

19. McGeoch, C.C., "Analyzing algorithms by simulation: variance reduction techniques and simulation speedups," *ACM Comp. Surveys* **24** (1992), 195–212.

20. Mishof, B., Anderson, C.L., and Hadrys, H., "A phylogeny of the damselfly genus *Calopteryx* (Odonata) using mitochondrial 16s rDNA markers," *Molec. Phylog. & Evol.* **15** (2000), 5–14.

21. Moret, B.M.E., "Towards a discipline of experimental algorithmics," *Proc. 5th & 6th DIMACS Implementation Challenges*, M. Goldwasser, D.S. Johnson, and C.C. McGeoch, eds, American Mathematical Society, Providence, 2003. Available at `www.cs.unm.edu/~moret/dimacs.ps`.

22. Olsen, G. J., Matsuda, H., Hagstrom, R., and Overbeek, R., "fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood," *Comput. Appl. Biosci.* **10** (1994): 41-48.

23. Penny, D., and Steel, M.A., "Distributions of tree comparison metrics—some new results," *Syst. Biol.* **42** (1993), 126–141.

24. Rambaut, A., and Grassly, N.C., "Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees," *Comp. Applic. Biosci.* **13** (1997), 235–238.

25. Ranwez, V. and Gascuel, O., "Quartet-based phylogenetic inference: improvements and limits," *Mol. Biol. Evol.* **18** (2001) (6):1103-1116.

26. Rodrigues-Trelles, F., Alarcon, L., and Fontdevila, A., "Molecular evolution and phylogeny of the *buzzatii* complex (*D. repleta* group): a maximum likelihood approach," *Mol. Biol. & Evol.* **17** (2000), 1112–1122.

27. Saitou, N., and Nei, M., "The neighbor-joining method: A new method for reconstructing phylogenetic trees," *Mol. Biol. & Evol.* **4** (1987), 406–425.

28. Strimmer, K., and von Haeseler, A., "Quartet puzzling: a maximum likelihood method for reconstructing tree topologies," *Mol. Biol. & Evol.* **13** (1996), 964–969.

29. Studier, J.A., and Keppler, K.J., "A note on the neighbor-joining method of Saitou and Nei," *Mol. Biol. & Evol.* **5** (1981), 729–731.

30. Szekeres, P.G. Muir, A.I., Spinage, L.D., Miller, J.E., Butler, S.I., Smith, A., Rennie, G.I., Murdock, P.R., Fitzgerald, L.R., Wu, H., McMillan, L.J., Guerrera, S., Vawter, L., Elshourbagy, N.A., Mooney, J.L., Bergsma, D.J., Wilson, S., and Chambers, J.K., "Neuromedin U is a potent agonist at the orphan G protein coupled receptor FM3," *J. Biol. Chem.* **275** (2000), 20247–20250.

31. Willson, S.J., "An error-correcting map for quartest can improve the signals for phylogenetic trees," *Mol. Biol. Evol.* **18**, 3 (2001), 344–351.